Parameter Learning with Hidden Variables

Davide Bacciu

Dipartimento di Informatica
Università di Pisa
bacciu@di.unipi.it

Apprendimento Automatico: Fondamenti - A.A. 2016/2017

Lecture Outline

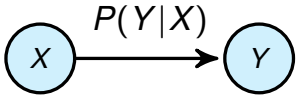
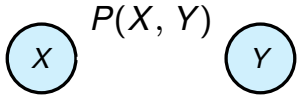
1. Learning with Hidden Variables
   - The Expectation Maximization (EM) Algorithm
   - Learning Gaussian Mixture Models
2. Concluding Remarks
3. Applications of Bayesian Learning
   - Semantic Analysis of Text
   - Image Understanding
   - Structure Learning

Part I

Learning with Hidden Variables in Bayesian Networks

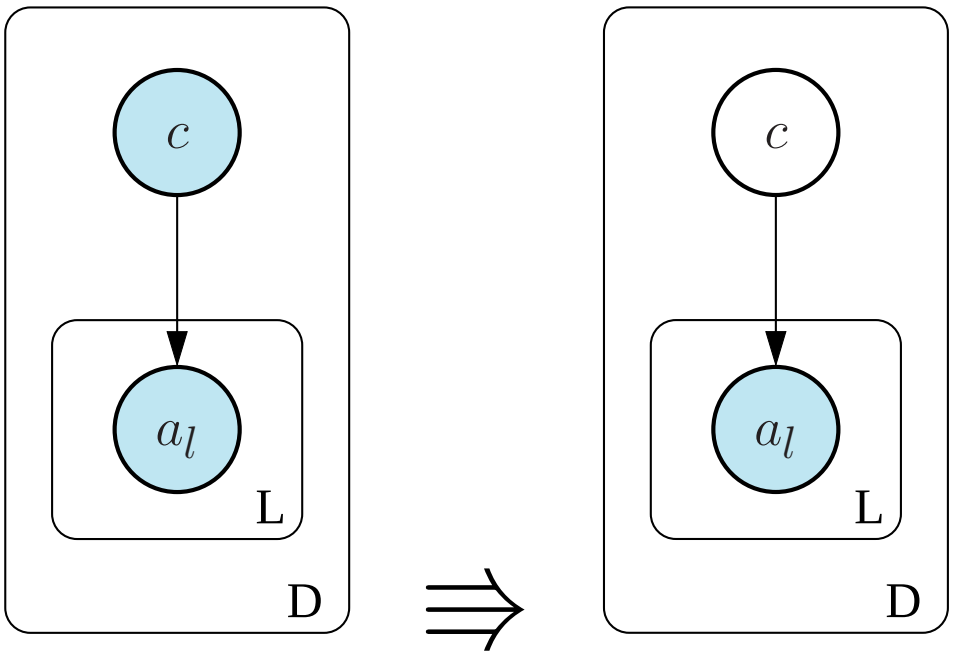
Learning with Bayesian Networks

Learning with Hidden Variables

		Structure	
		Fixed Structure 	Fixed Variables 
Data	Complete	Naive Bayes Calculate Frequencies (ML)	Discover dependencies from the data Structure Search Independence tests
	Incomplete	Latent variables EM Algorithm (ML) MCMC, VBEM (Bayesian)	Difficult Problem Structural EM
		Parameter Learning	Structure Learning

What if...

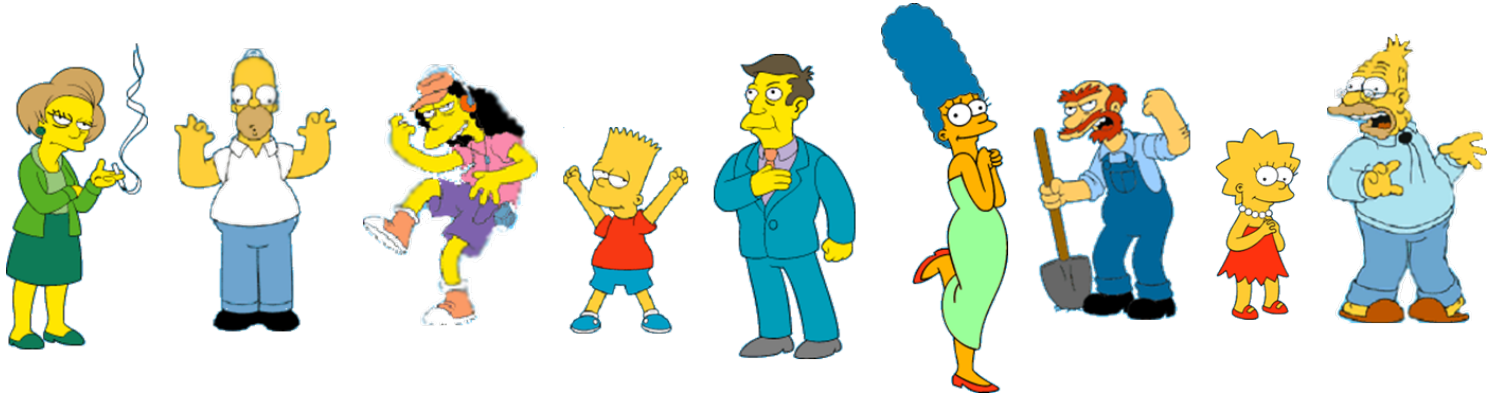
Sample classification is no longer observable



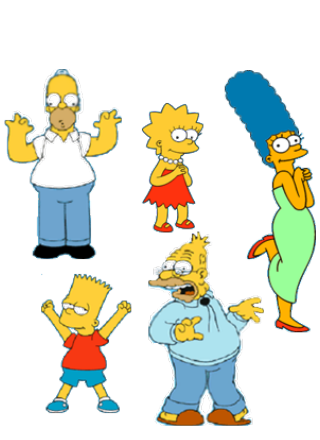
c is now an Hidden Variable within a (multinomial) Mixture Model

Mixture Model

No classification $\Rightarrow$ Seek a natural grouping



What is a natural grouping?



Family

School

Females

Males

Credit goes to Eamonn Keogh @ UCR

Dealing with Incomplete Data

- Incomplete data
  - Missing observations (imputation)
  - Unobserved random variables
- Key idea
  - Complete the data making hypothesis on the unobserved variables
  - Review the hypothesis
  - Iterate
- Expectation-Maximization Algorithm

  Expectation Compute the current estimate of the hidden variables (expected counts)

  Maximization Use the estimate to update the model parameters θ

Expectation-Maximization (EM) Algorithm

- Likelihood maximization procedure

$$\theta_{ML} = \arg\max_{\theta} \mathcal{L}(\theta|\mathbf{X}) = \arg\max_{\theta} P(\mathbf{X}|\theta)$$

- $\mathcal{L}(\theta|\mathbf{X})$ is the incomplete likelihood and $\mathbf{X}$ is the incomplete data containing only observed random variables
- Assume complete data $\mathbf{d} = (\mathbf{X}, \mathbf{Z})$ exists where $z \in \mathbf{Z}$ is an hidden or latent variable (e.g. c in the multinomial mixture example)
- Compute the complete likelihood

$$\mathcal{L}_c(\theta|\mathbf{d}) = P(\mathbf{d}|\theta) = P(\mathbf{X}, \mathbf{Z}|\theta) = P(\mathbf{Z}|\mathbf{X}, \theta)P(\mathbf{X}|\theta)$$

Key Idea

Maximize the complete likelihood $\mathcal{L}_c$ instead of $\mathcal{L}$

EM at a Glance

1. Initialize parameters $\theta^{(1)}$
2. Expectation Step (iteration k)

$$Q(\theta|\theta^{(k)}) = E_{\mathbf{Z}|\mathbf{X},\theta^{(k)}}[\log \mathcal{L}_c(\theta|\mathbf{X},\mathbf{Z})]$$

3. Maximization Step (iteration k)

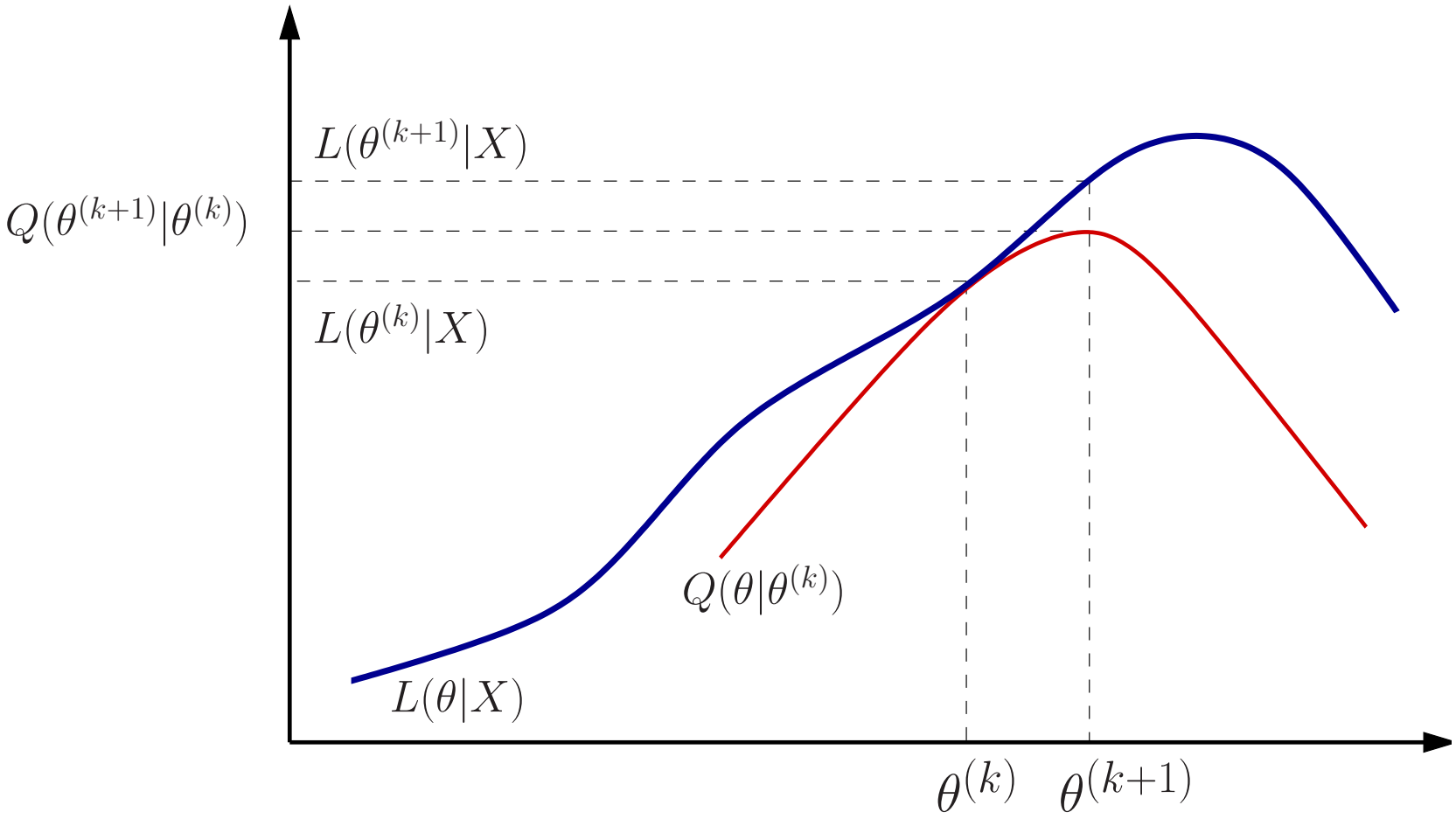
$$\theta^{(k+1)} = \arg\max_{\theta} Q(\theta|\theta^{(k)})$$

4. Iterate steps 2 and 3 with $k = k + 1$ until $\log \mathcal{L}_c$ stops increasing

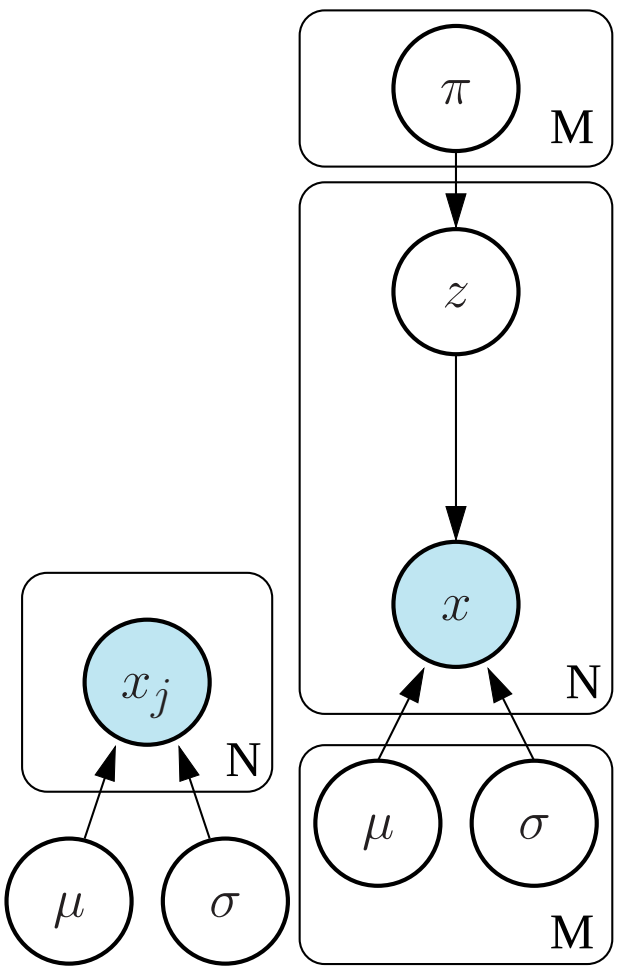
Theorem (Dempster, Laird and Rubin (1997))

An EM process ensures $\mathcal{L}(\theta^{(k+1)}|\mathbf{X}) \geq \mathcal{L}(\theta^{(k)}|\mathbf{X})$

EM Graphically



Learning Gaussian Mixture Models



- Previously - ML estimation for a single univariate Gaussian
- Now - EM estimation for M univariate Gaussians
  - Gaussian Mixture Model (GMM)
  - z_{jm} is the hidden mixture selection variable with prior $P(z_{jm}) = \pi_m$
  - Each observation x_j is generated by a single Gaussian $P(x_j|z_{jm}) \sim \mathcal{N}(\mu_m, \sigma_m)$
- Model parameters $\theta = (\pi, \mu, \sigma)$

GMM Likelihood

A set of i.i.d observations $\mathbf{X} = \{x_1, \ldots, x_n\}$ has incomplete likelihood

$$\mathcal{L}(\theta|\mathbf{X}) = P(\mathbf{X}|\theta) = \prod_{j=1}^{N} P(x_j|\theta)$$

Use marginalization to introduce the hidden variable z_{jm}

$$\mathcal{L}(\theta|\mathbf{X}) = \prod_{j=1}^{N} \sum_{m=1}^{M} P(x_j, z_{jm}|\theta)$$

Following the independence relationships in the graphical model this rewrites

$$\mathcal{L}(\theta|\mathbf{X}) = \prod_{j=1}^{N} \sum_{m=1}^{M} P(z_{jm}|\pi) P(x_j|z_{jm}, \mu, \sigma)$$

GMM Complete Likelihood

Assume we know the hidden assignment $z_{jm} = 1$ for each sample j, we have the following complete likelihood

$$\mathcal{L}_c(\theta|\mathbf{X},\mathbf{Z}) = \prod_{j=1}^{N}\sum_{m=1}^{M} z_{jm}\pi_m P(x_j|\mu_m,\sigma_m)$$

Can be better handled by taking the log

$$\log \mathcal{L}_c(\theta|\mathbf{X},\mathbf{Z}) = \sum_{j=1}^{N}\sum_{m=1}^{M} z_{jm}\log(\pi_m P(x_j|\mu_m,\sigma_m))$$

Now we can compute

$$\begin{aligned} Q(\theta|\theta^{(k)}) &= E_{\mathbf{Z}|\mathbf{X},\theta^{(k)}}[\log \mathcal{L}_c(\theta|\mathbf{X},\mathbf{Z})] \\ &= \sum_{j=1}^{N}\sum_{m=1}^{M} P(z_{jm}|x_j,\theta^{(k)})\log(\pi_m^{(k)} P(x_j|\mu_m^{(k)},\sigma_m^{(k)})) \end{aligned}$$

Expectation Step

To compute the auxiliary function

$$Q(\theta|\theta^{(k)}) = \sum_{j=1}^{N}\sum_{m=1}^{M} P(z_{jm}|x_j,\theta^{(k)}) \log(\pi_m^{(k)} P(x_j|\mu_m^{(k)},\sigma_m^{(k)}))$$

we only need to estimate the posterior $P(z_{jm}|x_j,\theta^{(k)})$ based on the current $\theta^{(k)}$ values

$$P(z_{jm}|x_j,\theta^{(k)}) = \frac{\pi_m^{(k)} P(x_j|\mu_m^{(k)},\sigma_m^{(k)})}{\sum_{m'} \pi_{m'}^{(k)} P(x_j|\mu_{m'}^{(k)},\sigma_{m'}^{(k)})}$$

Maximization Step

Maximization of the auxiliary function with respect to the parameters $\theta = (\pi, \mu, \sigma)$

$$Q(\theta|\theta^{(k)}) = \sum_{j=1}^{N}\sum_{m=1}^{M} P(z_{jm}|x_j, \theta^{(k)}) \log \pi_m^{(k)} + \sum_{j=1}^{N}\sum_{m=1}^{M} P(z_{jm}|x_j, \theta^{(k)}) \log P(x_j|\mu_m^{(k)}, \sigma_m^{(k)})$$

As usual this is performed by solving

$$\frac{\partial Q(\theta|\theta^{(k)})}{\partial \theta} = 0$$

while taking into account the sum-to-one constraint $\sum_{m=1}^{M} \pi_m = 1$

M-Step Learning Equations

Solving the independent maximization problems with respect to π_m, μ_m and σ_m yields

$$\pi_m^{(k+1)} = \frac{\sum_{j=1}^{N} P(z_{jm}|x_j, \theta^{(k)})}{N}$$

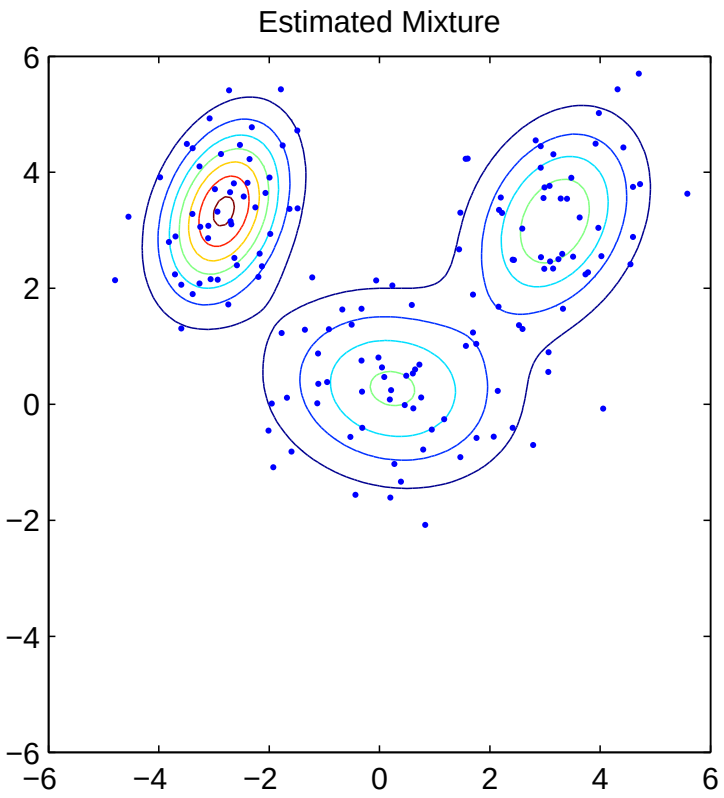
$$\mu_m^{(k+1)} = \frac{\sum_{j=1}^{N} x_j P(z_{jm}|x_j, \theta^{(k)})}{\sum_{j=1}^{N} P(z_{jm}|x_j, \theta^{(k)})}$$

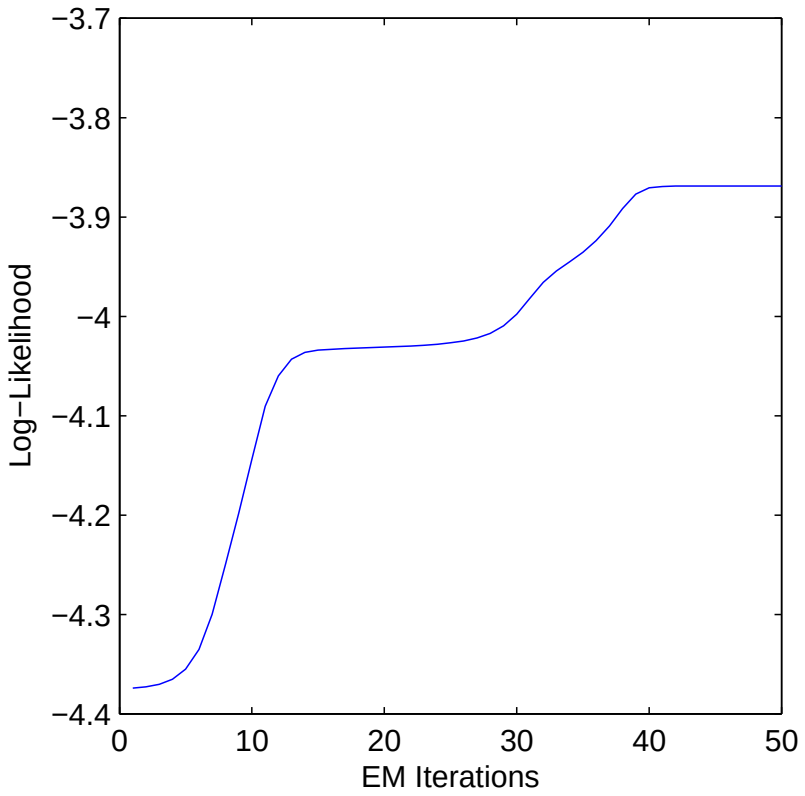
$$\sigma_m^{(k+1)} = \sqrt{\frac{\sum_{j=1}^{N} P(z_{jm}|x_j, \theta^{(k)})(x_j - \mu^{(k)})^2}{\sum_{j=1}^{N} P(z_{jm}|x_j, \theta^{(k)})}}$$

2D Gaussian Mixture Example

- Training data sampled from a number of Gaussian functions (Guess It!!) having common unit variance and different means
- 1500 test samples drawn from the same distribution for computing training-independent likelihood values (Overfitting!!)
- Task is to estimate a probability density for the data using a Gaussian Mixture Model
  - How many mixture components?

3 Gaussian Mixtures





The actual number of Gaussians generating the training data

Credit goes to Mark Girolami @ Glasgow University

The EM Algorithm in a Nut-Shell

The EM algorithm can be summarized by the optimization problem

$$\theta^{(k+1)} = \arg\max_{\theta} \sum_{\mathbf{z}} P(\mathbf{Z} = \mathbf{z}|\mathbf{X}, \theta^{(k)}) \log \mathcal{L}_c(\theta|\mathbf{X}, \mathbf{Z} = \mathbf{z})$$

The posterior $P(\mathbf{Z} = \mathbf{z}|\mathbf{X}, \theta^{(k)})$ provides an expected count for the event $\mathbf{z}$ that is the EM counterpart of basic ML counting

The expectation-maximization algorithm is an example of Unsupervised Learning

Take Home Messages on Parameter Learning

- Maximum Likelihood estimates provide straightforward solutions to both supervised and unsupervised learning tasks
- Maximum Likelihood is completely data-driven (handle with care!!)
  - Point estimates can degenerate if data is sparse
- Priors help regularizing the estimate (Smoothing)
- Bayesian learning might be the solution when dealing with little training data
- Expectation Maximization
  - Use it for dealing with unobserved data
  - Complete observations with hidden variables
  - Estimate unobserved counts
  - Compute observed terms using such pseudo-counts

Bayesian Learning in a Single Slide

- Bayesian learning is a powerful all-in-one model for
  - Modeling your knowledge about the world
  - Inference
  - Learning
- Bayesian learning uses conditional independence to
  - Break-down the joint probability into smaller pieces
  - Describe the world
- Bayesian learning is generative
  - Describes how data is generated
  - Learns the joint probability
- Bayesian learning is (often) parametric
  - Need to make hypothesis on the data-generating distribution
  - Need to choose suitable priors

Discriminative Vs Generative Learning

- Two paradigms for learning statistical classifiers
- Generative classifiers learn a model of the joint probability $P(c, x)$ of the features and the corresponding class label
  - Provide a model of how data is generated
  - Use Bayes rule to predict using class posterior $P(c|x)$
  - Bayesian Learning
- Discriminative classifiers directly model the class posterior $P(c|x)$ or use a discriminant function
  - Directly target at optimizing classification accuracy
  - Best to solve a specific estimation problem rather than trying to estimate the general joint density (Vapnik)
  - Linear threshold unit, neural networks, support vector machines, decision trees, ...

Why (not) Using Bayesian Learning

- Cons
  - Discriminative approaches often lead to better classification/regression accuracies
  - Inference may require a significant computational cost
  - Need to correctly *guess* the data generating and prior distributions
- Pro
  - Elegant and solid mathematical basis: lots of tools for inference, probabilistic reasoning, learning, assessing the reliability of the acquired knowledge
  - Straightforward to incorporate prior knowledge: data distributions and structure
  - Powerful and sound techniques for dealing with complex unsupervised learning tasks: vision, natural language processing, ...

Things we haven't seen

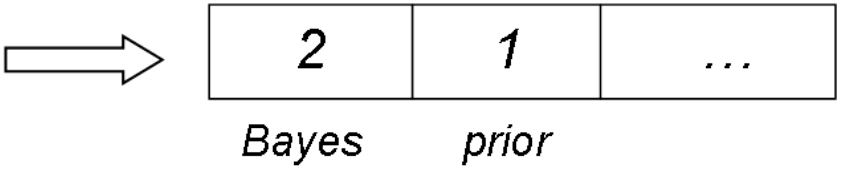
- Loads of them!!
- Fully Bayesian generative models
  - Treating maximum likelihood probabilities as random variables
  - Choosing appropriate priors
  - Variational and Montecarlo methods
- Learning the structure of a Bayesian Network
  - PC algorithm
  - Searching in the space of graphs structures
- Simultaneous learning of parameters and structure
  - Structural EM
- Advanced applications
  - Image and video understanding
  - Multimodal document analysis

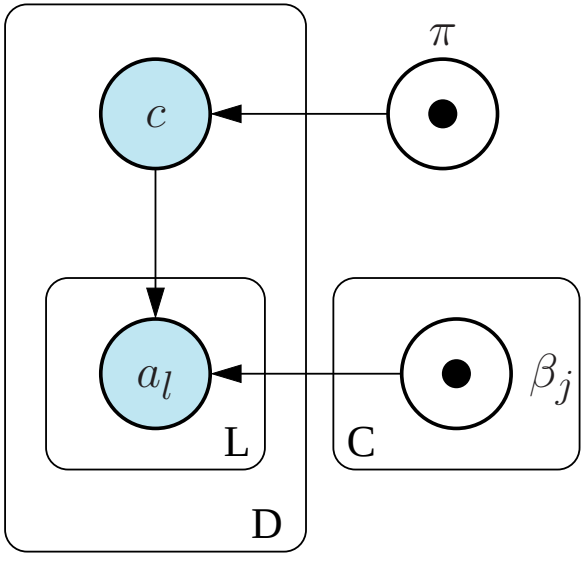
Part II

Applications of Bayesian Learning

Document Classification

The example shows that *the true hypothesis eventually dominates the Bayesian prediction.* This is characteristic of Bayesian learning. For any fixed prior that does not rule out…

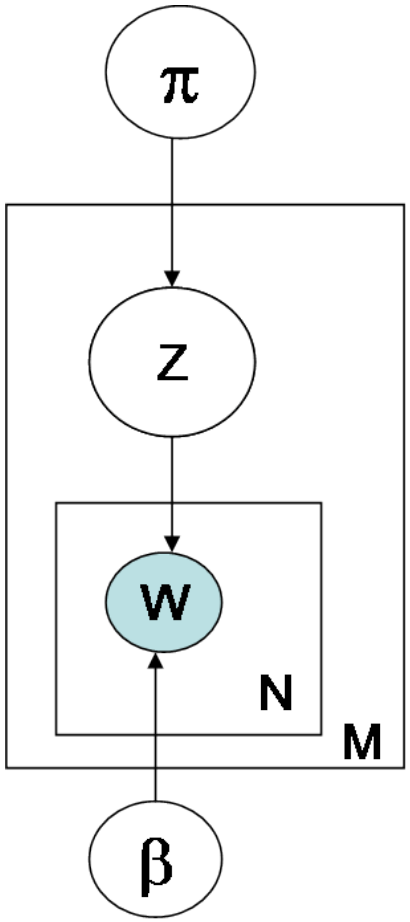




- Naive Bayes for classification of textual documents
- Need large annotated corpora
- What if we don't have such supervised information?

Generative Topic Models

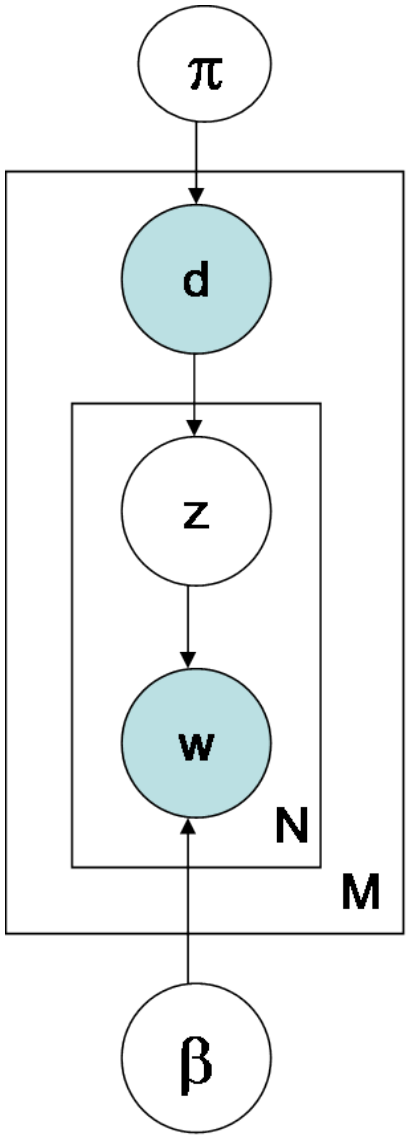
Unsupervised Naive Bayes



- Document classification becomes an unknown latent topic z
- Discovering the natural grouping of documents
- Expectation Maximization learning
- Unsupervised Naive Bayes allows choosing only one topic per document

The example shows that the true hypothesis eventually dominates the Bayesian prediction. This is characteristic of Bayesian learning. For any fixed prior that does not rule out…

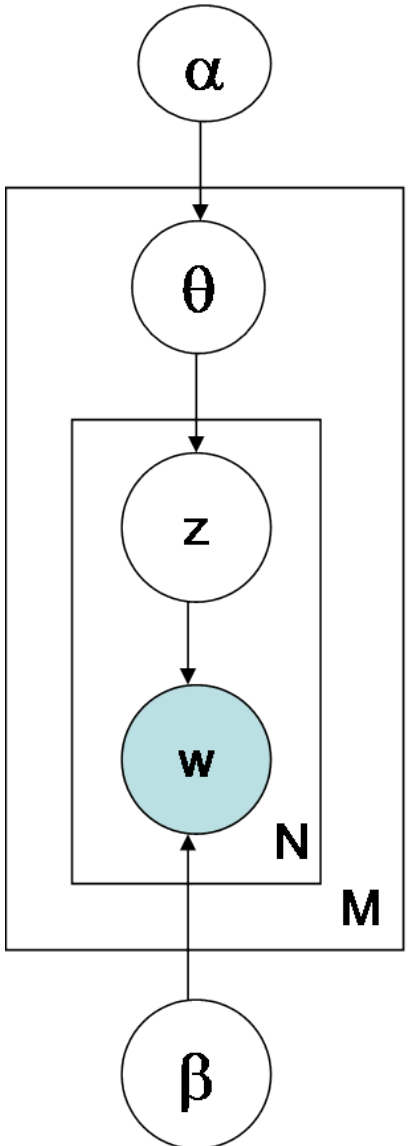
Generative Topic Models
Probabilistic Latent Semantic Analysis



- Choosing one topic per word
- Documents are mixtures of topics
- Not a fully Bayesian approach though
  - Generative model only for documents in the training set

The example shows that the true hypothesis eventually dominates the Bayesian prediction. This is characteristic of Bayesian learning. For any fixed prior that does not rule out...

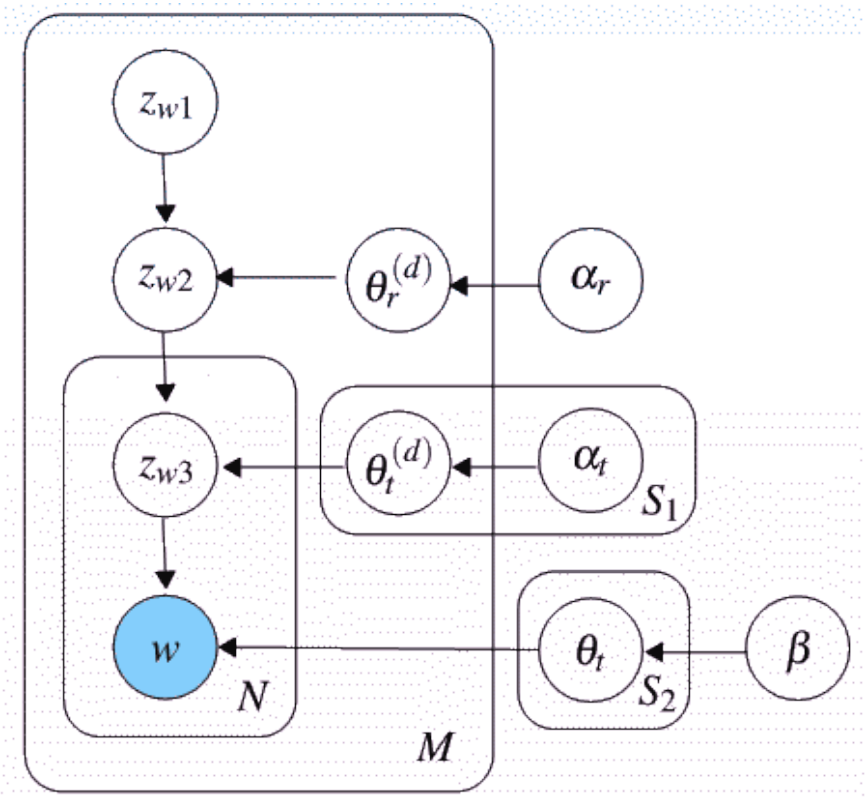
Generative Topic Models
Latent Dirichlet Allocation



- Document distribution becomes a random variable
  - Can generate documents outside the training set
- Distributed as a Dirichlet with hyperparameter α
- Parameter learning
  - Variational EM
  - Gibbs Sampling

Can Make it as Complex as You Wish!

Dependent Pachinko Allocation



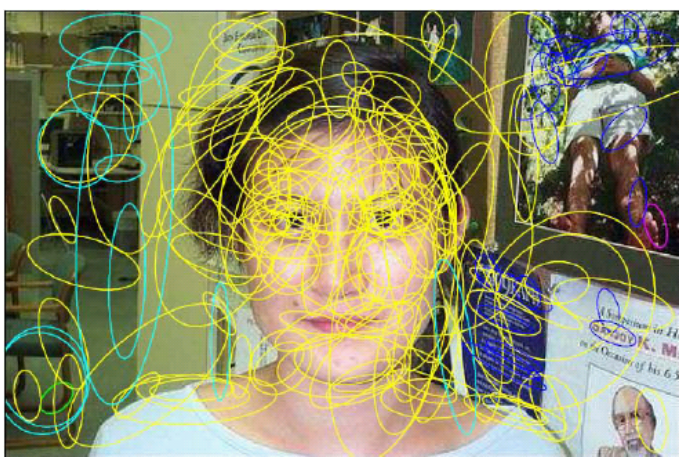
Hierarchical topic model

- Fixed hierarchy
- Fixed number of topics

Topic Models for Images

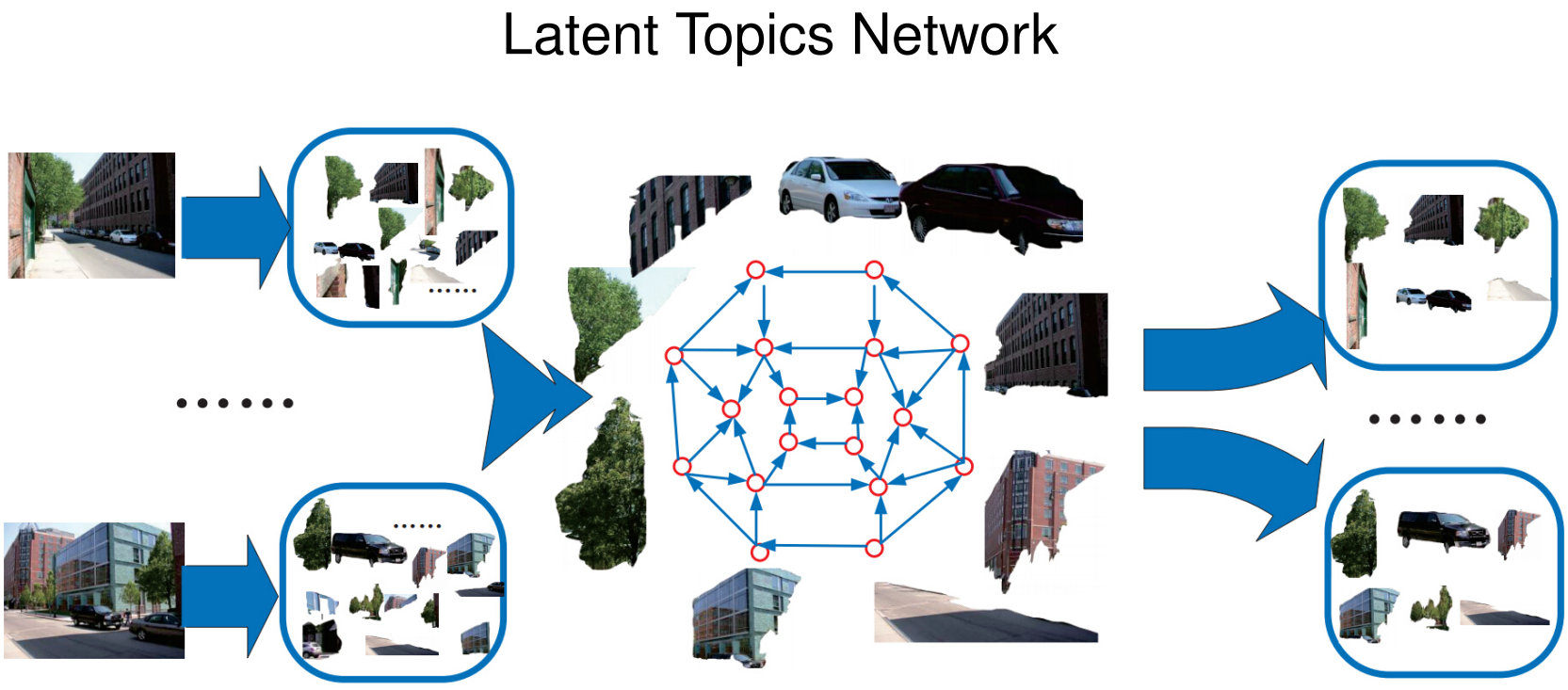
- Caltech-4 Dataset: cars, faces, motorbikes and airplanes on a cluttered background
- Fit a Latent Semantic Analysis model with 7 topics
- Determine the most likely topic of each visual word in image d_i

$$P(z_k|w_j, d_i) = \frac{P(z_k|d_i)P(w_j|z_k)}{\sum_{k=1}^{K} P(z_k|d_i)P(w_j|z_k)}$$



Sivic et al (2005) Discovering objects and their location in images. *Proc. of the 10th IEEE Int. Conf. on Computer Vision (ICCV'05)*

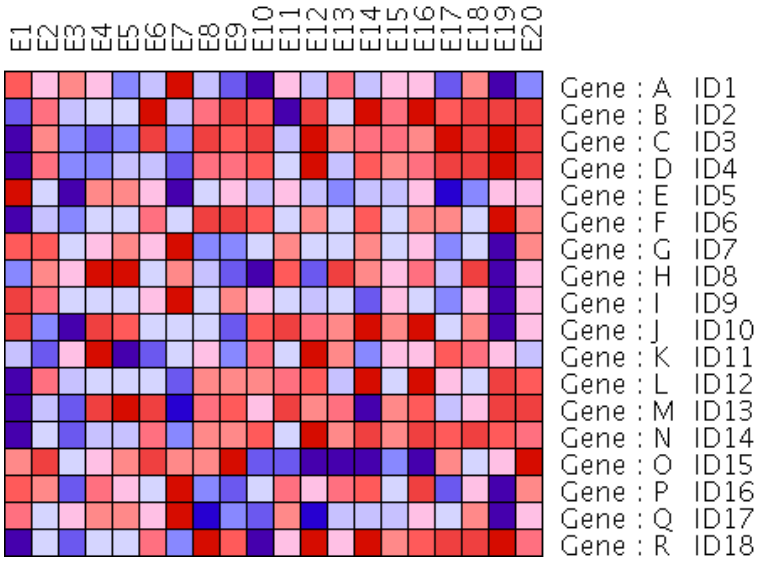
Learning to Segment Image Parts

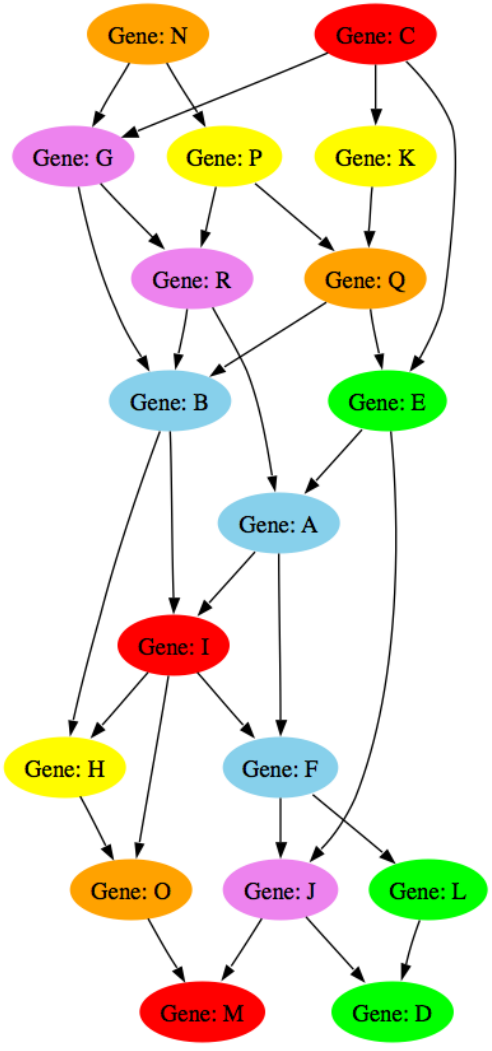


Yuan et al, A novel topic-level random walk framework for scene image co-segmentation, ECCV 2014

Discovering Gene Interaction Networks

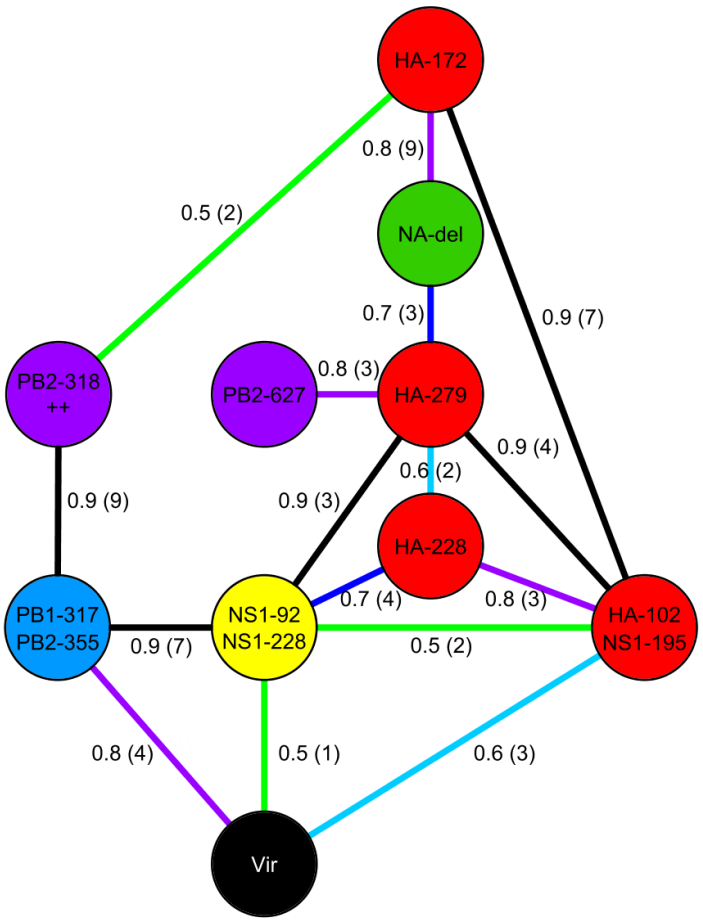
Gene Expression Data


E1 E2 E3 E4 E5 E6 E7 E8 E9 E10 E11 E12 E13 E14 E15 E16 E17 E18 E19 E20
Gene : A ID1
Gene : B ID2
Gene : C ID3
Gene : D ID4
Gene : E ID5
Gene : F ID6
Gene : G ID7
Gene : H ID8
Gene : I ID9
Gene : J ID10
Gene : K ID11
Gene : L ID12
Gene : M ID13
Gene : N ID14
Gene : O ID15
Gene : P ID16
Gene : Q ID17
Gene : R ID18



Assessing Influenza Virulence

Inferred Gene Bayesian Graphical Model for H5N1 Virulence



http://hiv.bio.ed.ac.uk/samLycett/research.html